Simulation - Lectures - Part I

Julien Berestycki -(adapted from François Caron's slides)

Part A Simulation and Statistical Programming

Hilary Term 2019

Simulation and Statistical Programming

- **Lectures on Simulation** (Prof. J. Berestycki):
  Tuesdays 2-3pm Weeks 1-8.
  LG.01.
- **Computer Lab on Statistical Programming** (Prof. G. Nicholls):
  Tuesday 3-5pm Weeks 1,2,3 Friday 9-11am Weeks 5,6,8.
  LG.02.
- **Departmental problem classes**:
  Mon. 2-3pm (Berestycki) or Thursday (Caterini). 1.30-2.30 pm - Weeks 3,5,7,TT1.
  LG.??,.
- Hand in problem sheet solutions by Thursday 10 am of previous week for both classes.
- Webpage:
  http://www.stats.ox.ac.uk/~berestyc/teaching/A12.html
- This course builds upon the notes and slides of Geoff Nicholls, Arnaud Doucet, Yee Whye Teh and Matti Vihola.

Outline

Introduction

Monte Carlo integration

Random variable generation

- Inversion Method
- Transformation Methods
- Rejection Sampling

Outline

Introduction

Monte Carlo integration

Random variable generation
- Inversion Method
- Transformation Methods
- Rejection Sampling

Monte Carlo Simulation Methods

- Computational tools for the simulation of random variables and the approximation of integrals/expectations.
- These simulation methods, aka Monte Carlo methods, are used in many fields including statistical physics, computational chemistry, statistical inference, genetics, finance etc.
- The Metropolis algorithm was named the top algorithm of the 20th century by a committee of mathematicians, computer scientists & physicists.
- With the dramatic increase of computational power, Monte Carlo methods are increasingly used.

Objectives of the Course

- Introduce the main tools for the simulation of random variables and the approximation of multidimensional integrals:
  - Integration by Monte Carlo,
  - inversion method,
  - transformation method,
  - rejection sampling,
  - importance sampling,
  - Markov chain Monte Carlo including Metropolis-Hastings.
- Understand the theoretical foundations and convergence properties of these methods.
- Learn to derive and implement specific algorithms for given random variables.

Computing Expectations

- Let X be either
  - a discrete random variable (r.v.) taking values in a countable or finite set Ω, with p.m.f. f_X
  - or a continuous r.v. taking values in $\Omega = \mathbb{R}^d$, with p.d.f. f_X
- Assume you are interested in computing

$$\begin{aligned} \theta &= \mathbb{E}\left(\phi(X)\right) \\ &= \begin{cases} \sum_{x\in\Omega} \phi(x) f_X(x) & \text{if } X \text{ is discrete} \\ \int_\Omega \phi(x) f_X(x) dx & \text{if } X \text{ is continuous} \end{cases} \end{aligned}$$

  where $\phi : \Omega \to \mathbb{R}$.
- It is impossible to compute θ exactly in most realistic applications.
- Even if it is possible (for Ω finite) the number of elements may be so huge that it is practically impossible
- Example: $\Omega = \mathbb{R}^d$, $X \sim \mathcal{N}(\mu, \Sigma)$ and $\phi(x) = \mathbb{I}\left(\sum_{k=1}^d x_k^2 \geq \alpha\right)$.
- Example: $\Omega = \mathbb{R}^d$, $X \sim \mathcal{N}(\mu, \Sigma)$ and $\phi(x) = \mathbb{I}\left(x_1 < 0, ..., x_d < 0\right)$.

Example: Queuing Systems

- Customers arrive at a shop and queue to be served. Their requests require varying amount of time.
- The manager cares about customer satisfaction and not excessively exceeding the 9am-5pm working day of his employees.
- Mathematically we could set up stochastic models for the arrival process of customers and for the service time based on past experience.
- **Question**: If the shop assistants continue to deal with all customers in the shop at 5pm, what is the probability that they will have served all the customers by 5.30pm?
- If we call $X \in \mathbb{N}$ the number of customers in the shop at 5.30pm then the probability of interest is

$$\mathbb{P}(X = 0) = \mathbb{E}(\mathbb{I}(X = 0)).$$

- For realistic models, we typically do not know analytically the distribution of X.

Example: Particle in a Random Medium

- A particle $(X_t)_{t=1,2,\ldots}$ evolves according to a stochastic model on $\Omega = \mathbb{R}^d$.
- At each time step t, it is absorbed with probability $1 - G(X_t)$ where $G : \Omega \to [0, 1]$.
- **Question**: What is the probability that the particle has not yet been absorbed at time T?
- The probability of interest is

$$\mathbb{P}\left(\text{not absorbed at time } T\right) = \mathbb{E}\left[G(X_1)G(X_2)\cdots G(X_T)\right].$$

- For realistic models, we cannot compute this probability.

Example: Ising Model

- The Ising model serves to model the behavior of a magnet and is the best known/most researched model in statistical physics.
- The magnetism of a material is modelled by the collective contribution of dipole moments of many atomic spins.
- Consider a simple 2D-Ising model on a finite lattice $\mathcal{G} = \{1, 2, ..., m\} \times \{1, 2, ..., m\}$ where each site $\sigma = (i, j)$ hosts a particle with a +1 or -1 spin modeled as a r.v. X_σ.
- The distribution of $X = \{X_\sigma\}_{\sigma \in \mathcal{G}}$ on $\{-1, 1\}^{m^2}$ is given by

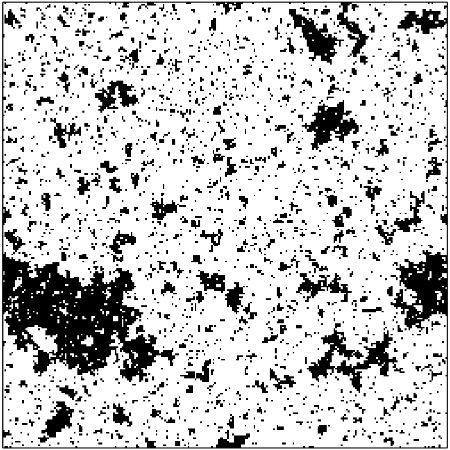
$$\pi(x) = \frac{\exp(-\beta U(x))}{Z_\beta}$$

where $\beta > 0$ is the inverse temperature and the potential energy is

$$U(x) = -J \sum_{\sigma \sim \sigma'} x_\sigma x_{\sigma'}$$

- Physicists are interested in computing $\mathbb{E}[U(X)]$ and Z_β.

Example: Ising Model



Sample from an Ising model for $m = 250$.

Bayesian Inference

- Suppose (X, Y) are both continuous r.v. with a joint density $f_{X,Y}(x, y)$.
- Think of Y as data, and X as unknown parameters of interest
- We have

$$f_{X,Y}(x, y) = f_X(x)\ f_{Y|X}(y|x)$$

  where, in many statistics problems, $f_X(x)$ can be thought of as a prior and $f_{Y|X}(y|x)$ as a likelihood function for a given $Y = y$.
- Using Bayes' rule, we have

$$f_{X|Y}(x|y) = \frac{f_X(x)\ f_{Y|X}(y|x)}{f_Y(y)}.$$

- For most problems of interest, $f_{X|Y}(x|y)$ does not admit an analytic expression and we cannot compute

$$\mathbb{E}\left(\phi(X)|Y = y\right) = \int \phi(x) f_{X|Y}(x|y) dx.$$

Outline

Monte Carlo Integration

Definition (Monte Carlo method)

Let X be either a discrete r.v. taking values in a countable or finite set Ω, with p.m.f. f_X, or a continuous r.v. taking values in $\Omega = \mathbb{R}^d$, with p.d.f. f_X. Consider

$$\theta = \mathbb{E}\left(\phi(X)\right) = \begin{cases} \sum_{x\in\Omega} \phi(x) f_X(x) & \text{if } X \text{ is discrete} \\ \int_\Omega \phi(x) f_X(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

where $\phi : \Omega \to \mathbb{R}$. Let $X_1, ..., X_n$ be i.i.d. r.v. with p.d.f. (or p.m.f.) f_X. Then

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i),$$

is called the Monte Carlo estimator of the expectation θ.

- Monte Carlo methods can be thought of as a stochastic way to approximate integrals.

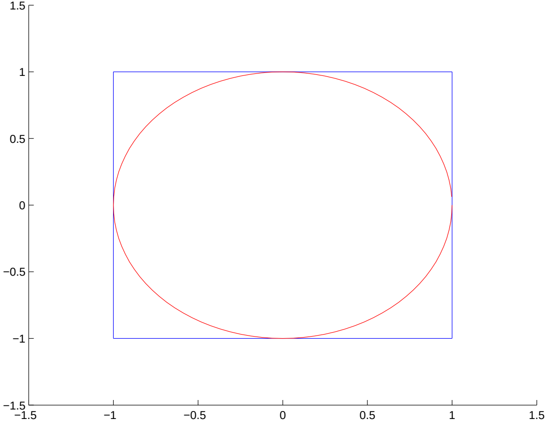
Monte Carlo Integration

Algorithm 1 Monte Carlo Algorithm

- Simulate independent $X_1, ..., X_n$ with p.m.f. or p.d.f. f_X
- Return $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i)$.

Computing Pi with Monte Carlo Methods

- Consider the 2×2 square, say $\mathcal{S} \subseteq \mathbb{R}^2$ with inscribed disk $\mathcal{D}$ of radius 1.



A 2×2 square $\mathcal{S}$ with inscribed disk $\mathcal{D}$ of radius 1.

Computing Pi with Monte Carlo Methods

- We have
$$\frac{\int\int_{\mathcal{D}} dx_1 dx_2}{\int\int_{\mathcal{S}} dx_1 dx_2} = \frac{\pi}{4}.$$

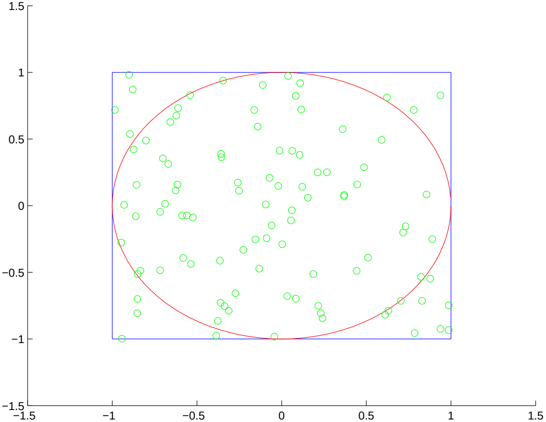
- How could you estimate this quantity through simulation?
$$\begin{aligned}\frac{\int\int_{\mathcal{D}} dx_1 dx_2}{\int\int_{\mathcal{S}} dx_1 dx_2} &= \int\int_{\mathcal{S}} \mathbb{I}\left((x_1, x_2) \in \mathcal{D}\right) \frac{1}{4} dx_1 dx_2 \\ &= \mathbb{E}\left[\phi(X_1, X_2)\right] = \theta\end{aligned}$$
where the expectation is w.r.t. the uniform distribution on $\mathcal{S}$ and
$$\phi(X_1, X_2) = \mathbb{I}\left((X_1, X_2) \in \mathcal{D}\right).$$

- To sample uniformly on $\mathcal{S} = (-1, 1) \times (-1, 1)$ then simply use
$$X_1 = 2U_1 - 1, \ X_2 = 2U_2 - 1$$
where $U_1, U_2 \sim \mathcal{U}(0, 1)$.

Computing Pi with Monte Carlo Methods

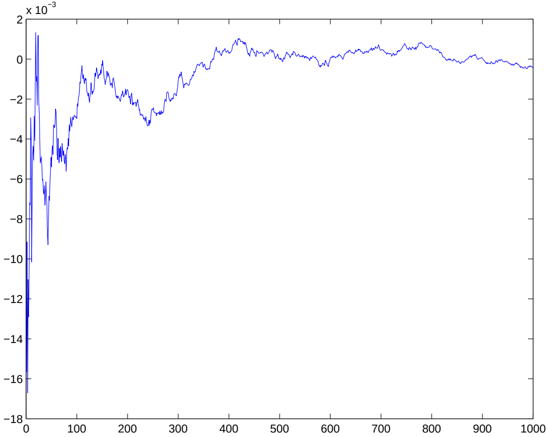
```r
n <- 1000
x <- array(0, c(2,1000))
t <- array(0, c(1,1000))

for (i in 1:1000) {
  # generate point in square
  x[1,i] <- 2*runif(1)-1
  x[2,i] <- 2*runif(1)-1

  # compute phi(x); test whether in disk
  if (x[1,i]*x[1,i] + x[2,i]*x[2,i] <= 1) {
    t[i] <- 1
  } else {
    t[i] <- 0
  }
}
print(sum(t)/n*4)
```

Computing Pi with Monte Carlo Methods



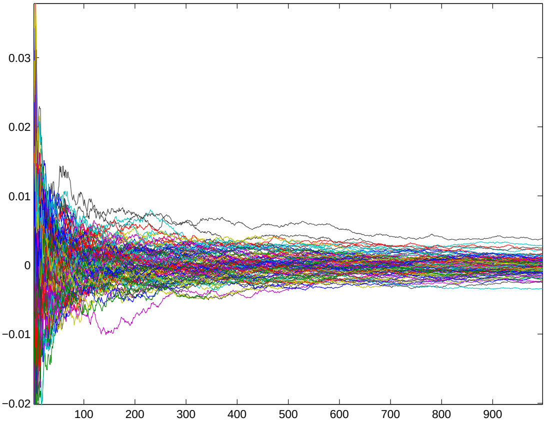
A 2×2 square $\mathcal{S}$ with inscribed disk $\mathcal{D}$ of radius 1 and Monte Carlo samples.

Computing Pi with Monte Carlo Methods



$\hat{\theta}_n - \theta$ as a function of the number of samples n.

Computing Pi with Monte Carlo Methods



$\hat{\theta}_n - \theta$ as a function of the number of samples n, 100 independent realizations.

Applications

- *Toy example*: simulate a large number n of independent r.v. $X_i \sim \mathcal{N}(\mu, \Sigma)$ and

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I} \left(\sum_{k=1}^{d} X_{k,i}^2 \geq \alpha \right).$$

- *Queuing*: simulate a large number n of days using your stochastic models for the arrival process of customers and for the service time and compute

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I} \left(X_i = 0 \right)$$

  where X_i is the number of customers in the shop at 5.30pm for ith sample.
- *Particle in Random Medium*: simulate a large number n of particle paths $(X_{1,i}, X_{2,i}, ..., X_{T,i})$ where $i = 1, ..., n$ and compute

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} G(X_{1,i}) G(X_{2,i}) \cdots G(X_{T,i})$$

Monte Carlo Integration: Properties

- **Proposition**: Assume $\theta = \mathbb{E}\left(\phi(X)\right)$ exists. Then the Monte Carlo estimator $\hat{\theta}_n$ has the following properties
  - Unbiasedness
  $$\mathbb{E}\left(\hat{\theta}_n\right) = \theta$$
  - Strong consistency
  $$\hat{\theta}_n \to \theta \text{ almost surely as } n \to \infty$$
- *Proof*: We have
$$\mathbb{E}\left(\hat{\theta}_n\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\phi(X_i)\right) = \theta.$$

  Strong consistency is a consequence of the strong law of large numbers applied to $Y_i = \phi(X_i)$ which is applicable as $\theta = \mathbb{E}\left(\phi(X)\right)$ is assumed to exist.

Monte Carlo Integration: Central Limit Theorem

- **Proposition**: Assume $\theta = \mathbb{E}\left(\phi(X)\right)$ and $\sigma^2 = \mathbb{V}\left(\phi(X)\right)$ exist then

$$\mathbb{E}\left((\hat{\theta}_n - \theta)^2\right) = \mathbb{V}\left(\hat{\theta}_n\right) = \frac{\sigma^2}{n}$$

  and

$$\frac{\sqrt{n}}{\sigma}\left(\hat{\theta}_n - \theta\right) \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

- Proof. We have $\mathbb{E}\left((\hat{\theta}_n - \theta)^2\right) = \mathbb{V}\left(\hat{\theta}_n\right)$ as $\mathbb{E}\left(\hat{\theta}_n\right) = \theta$ and

$$\mathbb{V}\left(\hat{\theta}_n\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\left(\phi(X_i)\right) = \frac{\sigma^2}{n}.$$

  The CLT applied to $Y_i = \phi(X_i)$ tells us that

$$\frac{Y_1 + \cdots + Y_n - n\theta}{\sigma\sqrt{n}} \xrightarrow{\text{d}} \mathcal{N}(0,1)$$

  so the result follows as $\hat{\theta}_n = \frac{1}{n}\left(Y_1 + \cdots + Y_n\right)$.

Monte Carlo Integration: Variance Estimation

- **Proposition**: Assume $\sigma^2 = \mathbb{V}(\phi(X))$ exists then

$$S^2_{\phi(X)} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\phi(X_i) - \hat{\theta}_n\right)^2$$

  is an unbiased sample variance estimator of σ^2.
- Proof. Let $Y_i = \phi(X_i)$ then we have

$$\begin{aligned}
\mathbb{E}\left(S^2_{\phi(X)}\right) &= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\left(\left(Y_i - \overline{Y}\right)^2\right) \\
&= \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n} Y_i^2 \quad - n\overline{Y}^2\right) \\
&= \frac{n\left(\mathbb{V}(Y) + \theta^2\right) - n\left(\mathbb{V}\left(\overline{Y}\right) + \theta^2\right)}{n-1} \\
&= \mathbb{V}(Y) = \mathbb{V}(\phi(X)).
\end{aligned}$$

  where $Y = \phi(X)$ and $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

How Good is The Estimator?

- Chebyshev's inequality yields the bound

$$\mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| > c\frac{\sigma}{\sqrt{n}}\right) \leq \frac{\mathbb{V}\left(\hat{\theta}_n\right)}{c^2\sigma^2/n} = \frac{1}{c^2}.$$

- Another estimate follows from the CLT for large n

$$\frac{\sqrt{n}}{\sigma}\left(\hat{\theta}_n - \theta\right) \overset{d}{\approx} \mathcal{N}(0,1) \Rightarrow \mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| > c\frac{\sigma}{\sqrt{n}}\right) \approx 2\left(1 - \Phi(c)\right).$$

- Hence by choosing $c = c_\alpha$ s.t. $2\left(1 - \Phi(c_\alpha)\right) = \alpha$, an approximate $(1-\alpha)100\%$-CI for θ is

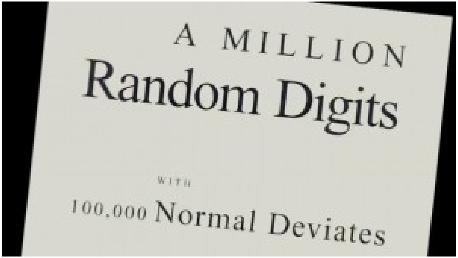
$$\left(\hat{\theta}_n \pm c_\alpha \frac{\sigma}{\sqrt{n}}\right) \approx \left(\hat{\theta}_n \pm c_\alpha \frac{S_{\phi(X)}}{\sqrt{n}}\right).$$

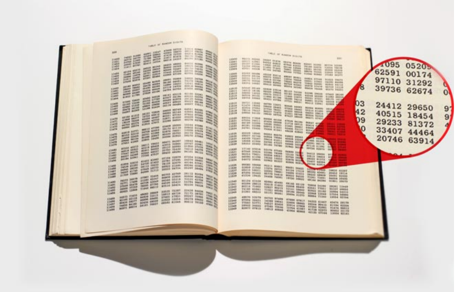
Monte Carlo Integration

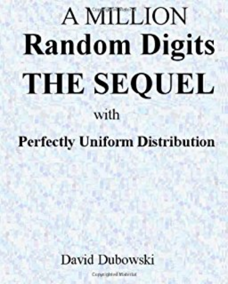
- Whatever being Ω; e.g. $\Omega = \mathbb{R}$ or $\Omega = \mathbb{R}^{1000}$, the error is still in $\sigma/\sqrt{n}$.
- This is in contrast with deterministic methods. The error in a product trapezoidal rule in d dimensions is $\mathcal{O}(n^{-2/d})$ for twice continuously differentiable integrands.
- It is sometimes said erroneously that it beats the curse of dimensionality but this is generally not true as σ^2 typically depends of $\dim(\Omega)$.

Pseudo-random numbers

- The aim of the game is to be able to generate complicated random variables and stochastic models.
- Henceforth, we will assume that we have access to a sequence of independent random variables $(U_i, i \geq 1)$ that are uniformly distributed on $(0,1)$; i.e. $U_i \sim \mathcal{U}[0,1]$.
- In R, the command `u←runif(100)` return 100 realizations of uniform r.v. in $(0,1)$.
- Strictly speaking, we only have access to pseudo-random (deterministic) numbers.
- The behaviour of modern random number generators (constructed on number theory) resembles mathematical random numbers in many respects: standard statistical tests for uniformity, independence, etc. do not show significant deviations.







> *If you like this book, I highly recommend that you read it in the original binary. As with most translations, conversion from binary to decimal frequently causes a loss of information and, unfortunately, it's the most significant digits that are lost in the conversion.*

Or this somewhat more subtle nerd-joke, by BJ from Waltford, England:

> *For a supposedly serious reference work the omission of an index is a major impediment. I hope this will be corrected in the next edition.*

...or from Fuat C. Baran:

> *A great read. Captivating. I couldn't put it down. I would have given it five stars, but sadly there were too many distracting typos. For example: 46453 13987. Hopefully they will correct them in the next edition.*

Or D.C. Froemke's one-star review:

Outline

Introduction

Monte Carlo integration

Random variable generation
- Inversion Method
- Transformation Methods
- Rejection Sampling

Outline

Introduction

Monte Carlo integration

Random variable generation

- Inversion Method
- Transformation Methods
- Rejection Sampling

Generating Random Variables Using Inversion

- A function $F : \mathbb{R} \to [0, 1]$ is a cumulative distribution function (cdf) if
  - F is increasing; i.e. if $x \leq y$ then $F(x) \leq F(y)$
  - F is right continuous; i.e. $F(x + \epsilon) \to F(x)$ as $\epsilon \to 0$ ($\epsilon > 0$)
  - $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to +\infty$.
- A random variable $X \in \mathbb{R}$ has cdf F if $\mathbb{P}(X \leq x) = F(x)$ for all $x \in \mathbb{R}$.
- If F is differentiable on $\mathbb{R}$, with derivative f, then X is continuously distributed with probability density function (pdf) f.

Generating Random Variables Using Inversion

- **Proposition**. Let F be a continuous and strictly increasing cdf on $\mathbb{R}$, with inverse $F^{-1}:[0,1]\rightarrow\mathbb{R}$. Let $U\sim\mathcal{U}[0,1]$ then $X=F^{-1}(U)$ has cdf F.
- Proof. We have

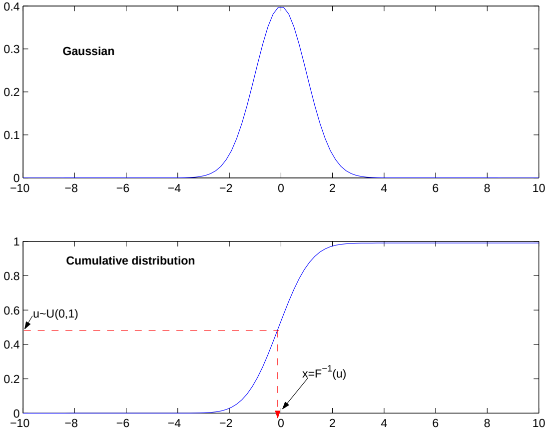
$$\begin{aligned}\mathbb{P}(X\leq x) &= \mathbb{P}\left(F^{-1}(U)\leq x\right)\\ &= \mathbb{P}(U\leq F(x))\\ &= F(x).\end{aligned}$$

- **Proposition**. Let F be a cdf on $\mathbb{R}$ and define its generalized inverse $F^{-1}:[0,1]\rightarrow\mathbb{R}$,

$$F^{-1}(u)=\inf\{x\in\mathbb{R};F(x)\geq u\}.$$

Let $U\sim\mathcal{U}[0,1]$ then $X=F^{-1}(U)$ has cdf F.

Illustration of the Inversion Method



Top: pdf of a Gaussian r.v., bottom: associated cdf.

Examples

- *Weibull distribution*. Let $\alpha, \lambda > 0$ then the Weibull cdf is given by

$$F(x) = 1 - \exp\left(-\lambda x^{\alpha}\right), \; x \geq 0.$$

We calculate

$$\begin{aligned} u &= F(x) \Leftrightarrow \log\left(1-u\right) = -\lambda x^{\alpha} \\ &\Leftrightarrow x = \left(-\frac{\log\left(1-u\right)}{\lambda}\right)^{1/\alpha}. \end{aligned}$$

- As $(1-U) \sim \mathcal{U}[0,1]$ when $U \sim \mathcal{U}[0,1]$ we can use

$$X = \left(-\frac{\log U}{\lambda}\right)^{1/\alpha}.$$

Examples

- *Cauchy distribution*. It has pdf and cdf
$$f(x) = \frac{1}{\pi\left(1+x^2\right)},\ F(x) = \frac{1}{2} + \frac{arc\tan x}{\pi}$$
We have
$$\begin{aligned} u &= F(x) \Leftrightarrow u = \frac{1}{2} + \frac{arc\tan x}{\pi} \\ &\Leftrightarrow x = \tan\left(\pi\left(u - \frac{1}{2}\right)\right) \end{aligned}$$

- *Logistic distribution.* It has pdf and cdf
$$\begin{aligned} f(x) &= \frac{\exp(-x)}{\left(1+\exp(-x)\right)^2},\ F(x) = \frac{1}{1+\exp(-x)} \\ &\Leftrightarrow x = \log\left(\frac{u}{1-u}\right). \end{aligned}$$

- Practice: Derive an algorithm to simulate from an Exponential random variable with rate $\lambda > 0$.

Generating Discrete Random Variables Using Inversion

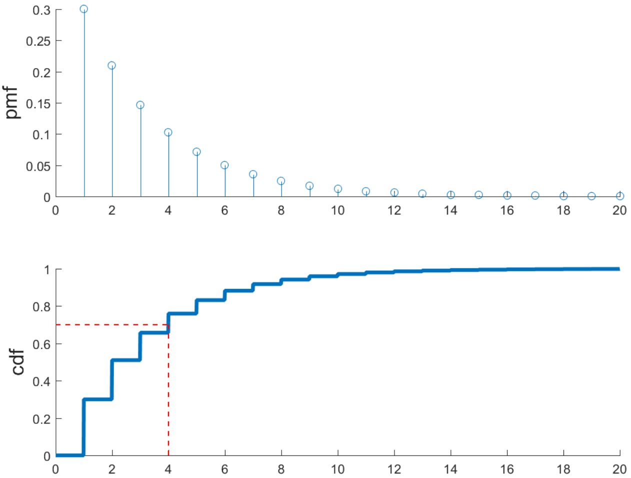
- If X is a discrete $\mathbb{N}-$r.v. with $\mathbb{P}(X = n) = p(n)$, we get $F(x) = \sum_{j=0}^{\lfloor x \rfloor} p(j)$ and $F^{-1}(u)$ is $x \in \mathbb{N}$ such that

$$\sum_{j=0}^{x-1} p(j) < u \leq \sum_{j=0}^{x} p(j)$$

with the LHS$= 0$ if $x = 0$.

- Note: the mapping at the values $F(n)$ are irrelevant.
- Note: the same method is applicable to any discrete valued r.v. X, $\mathbb{P}(X = x_n) = p(n)$.

Illustration of the Inversion Method: Discrete case



Example: Geometric Distribution

- If $0 < p < 1$ and $q = 1 - p$ and we want to simulate $X \sim \text{Geom}(p)$ then

$$p(x) = pq^{x-1}, F(x) = 1 - q^x \qquad x = 1, 2, 3...$$

- The smallest $x \in \mathbb{N}$ giving $F(x) \geq u$ is the smallest $x \geq 1$ satisfying

$$x \geq \log(1-u)/\log(q)$$

and this is given by

$$x = F^{-1}(u) = \left\lceil \frac{\log(1-u)}{\log(q)} \right\rceil$$

where $\lceil x \rceil$ rounds up and we could replace $1 - u$ with u.

Outline

Transformation Methods

- Suppose we have a random variable $Y \sim Q$, $Y \in \Omega_Q$, which we can simulate (eg, by inversion) and some other variable $X \sim P$, $X \in \Omega_P$, which we wish to simulate.
- Suppose we can find a function $\varphi : \Omega_Q \to \Omega_P$ with the property that $X = \varphi(Y)$.
- Then we can simulate from X by first simulating $Y \sim Q$, and then set $X = \varphi(Y)$.
- Inversion is a special case of this idea.
- We may generalize this idea to take functions of collections of variables with different distributions.

Transformation Methods

- Example: Let Y_i, $i = 1, 2, ..., \alpha$, be iid variables with $Y_i \sim \text{Exp}(1)$ and $X = \beta^{-1} \sum_{i=1}^{\alpha} Y_i$ then $X \sim \text{Gamma}(\alpha, \beta)$.
  Proof: The MGF of the random variable X is

$$\mathbb{E}\left(e^{tX}\right) = \prod_{i=1}^{\alpha} \mathbb{E}\left(e^{\beta^{-1} t Y_i}\right) = (1 - t/\beta)^{-\alpha}$$

  which is the MGF of a $\text{Gamma}(\alpha, \beta)$ variate.
  Incidentally, the $\text{Gamma}(\alpha, \beta)$ density is $f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ for $x > 0$.

- Practice: A generalized gamma variable Z with parameters $a > 0, b > 0, \sigma > 0$ has density

$$f_Z(z) = \frac{\sigma b^a}{\Gamma(a/\sigma)} z^{a-1} e^{-(bz)^\sigma}.$$

  Derive an algorithm to simulate from Z.

Transformation Methods: Box-Muller Algorithm

- For continuous random variables, a tool is the transformation/change of variables formula for pdf.
- **Proposition**. If $R^2 \sim \text{Exp}(\frac{1}{2})$ and $\Theta \sim \mathcal{U}[0, 2\pi]$ are independent then $X = R\cos\Theta$, $Y = R\sin\Theta$ are independent with $X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(0,1)$.
  Proof: We have $f_{R^2,\Theta}(r^2, \theta) = \frac{1}{2}\exp\left(-r^2/2\right)\frac{1}{2\pi}$ and

$$f_{X,Y}(x,y) = f_{R^2,\Theta}(r^2,\theta)\left|\det\frac{\partial(r^2,\theta)}{\partial(x,y)}\right|$$

  where

$$\left|\det\frac{\partial(r^2,\theta)}{\partial(x,y)}\right|^{-1} = \left|\det\begin{pmatrix} \frac{\partial x}{\partial r^2} & \frac{\partial x}{\partial\theta} \\ \frac{\partial y}{\partial r^2} & \frac{\partial y}{\partial\theta} \end{pmatrix}\right| = \left|\det\begin{pmatrix} \frac{\cos\theta}{2r} & -r\sin\theta \\ \frac{\sin\theta}{2r} & r\cos\theta \end{pmatrix}\right| = \frac{1}{2}.$$

Transformation Methods: Box-Muller Algorithm

- Let $U_1 \sim \mathcal{U}[0,1]$ and $U_2 \sim \mathcal{U}[0,1]$ then

$$\begin{aligned} R^2 &= -2\log(U_1) \sim \mathrm{Exp}\left(\frac{1}{2}\right) \\ \Theta &= 2\pi U_2 \sim \mathcal{U}[0,2\pi] \end{aligned}$$

  and

$$\begin{aligned} X &= R\cos\Theta \sim \mathcal{N}(0,1) \\ Y &= R\sin\Theta \sim \mathcal{N}(0,1), \end{aligned}$$

- This still requires evaluating $\log, \cos$ and $\sin$.

Simulating Multivariate Normal

- Let consider $X \in \mathbb{R}^d$, $X \sim N(\mu, \Sigma)$ where μ is the mean and Σ is the (positive definite) covariance matrix.

$$f_X(x) = (2\pi)^{-d/2} |\det \Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right).$$

- **Proposition**. Let $Z = (Z_1, ..., Z_d)$ be a collection of d independent standard normal random variables. Let L be a real $d \times d$ matrix satisfying

$$LL^T = \Sigma,$$

and

$$X = LZ + \mu.$$

Then

$$X \sim \mathcal{N}(\mu, \Sigma).$$

Simulating Multivariate Normal

- Proof. We have $f_Z(z) = (2\pi)^{d/2} \exp\left(-\frac{1}{2} z^T z\right)$.The joint density of the new variables is

$$f_X(x) = f_Z(z) \left| \det \frac{\partial z}{\partial x} \right|$$

where $\frac{\partial z}{\partial x} = L^{-1}$ and $\det(L) = \det(L^T)$ so $\det(L^2) = \det(\Sigma)$, and $\det(L^{-1}) = 1/\det(L)$ so $\det(L^{-1}) = \det(\Sigma)^{-1/2}$. Also

$$\begin{aligned} z^T z &= (x-\mu)^T \left(L^{-1}\right)^T L^{-1} (x-\mu) \\ &= (x-\mu)^T \Sigma^{-1} (x-\mu). \end{aligned}$$

- If $\Sigma = VDV^T$ is the eigendecomposition of Σ, we can pick $L = VD^{1/2}$.
- Cholesky factorization $\Sigma = LL^T$ where L is a lower triangular matrix.
- See numerical analysis.

Outline

Rejection Sampling

- Let X be a continuous r.v. on Ω with pdf f_X
- Consider a continuous rv variable $U > 0$ such that the conditional pdf of U given $X = x$ is

$$f_{U|X}(u|x) = \begin{cases} \frac{1}{f_X(x)} & \text{if } u < f_X(x) \\ 0 & \text{otherwise} \end{cases}$$

- The joint pdf of (X, U) is

$$\begin{aligned} f_{X,U}(x,u) &= f_X(x) \times f_{U|X}(u|x) \\ &= f_X(x) \times \frac{1}{f_X(x)} \mathbb{I}(0 < u < f_X(x)) \\ &= \mathbb{I}(0 < u < f_X(x)) \end{aligned}$$

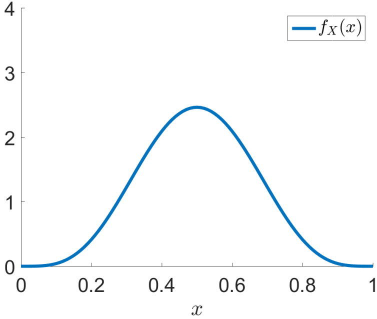
- Uniform distribution on the set $\mathcal{A} = \{(x,u) | 0 < u < f_X(x), x \in \Omega\}$

Rejection Sampling

Theorem (Fundamental Theorem of simulation)

Let X be a rv on Ω with pdf or pmf f_X. Simulating X is equivalent to simulating

$$(X, U) \sim \text{Unif}(\{(x, u) | x \in \Omega, 0 < u < f_X(x)\})$$



Rejection Sampling

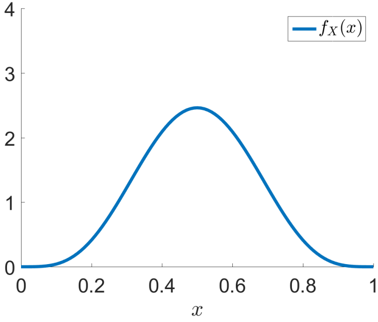
- Direct sampling of (X, U) uniformly over the set $\mathcal{A}$ is in general challenging
- Let $\mathcal{S} \supseteq \mathcal{A}$ be a bigger set such that simulating uniform rv on $\mathcal{S}$ is easy
- Rejection sampling technique:
  1. Simulate $(Y, V) \sim \mathrm{Unif}(\mathcal{S})$, with simulated values y and v
  2. if $(y, v) \in \mathcal{A}$ then stop and return $X = y$,$U = v$,
  3. otherwise go back to 1.
- The resulting rv (X, U) is uniformly distributed on $\mathcal{A}$
- X is marginally distributed from f_X

Example: Beta density

- Let $X \sim \mathrm{Beta}(5,5)$ be a continuous rv with pdf

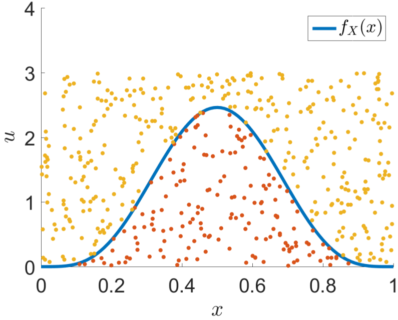
$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1},\ 0 < x < 1$$

  where $\alpha = \beta = 5$.
- $f_X(x)$ is upper bounded by 3 on $[0,1]$.



Example: Beta density

- Let $\mathcal{S} = \{(y, v) | y \in [0,1], v \in [0,3]\}$
  1. Simulate $Y \sim \mathcal{U}([0,1])$ and $V \sim \mathcal{U}([0,3])$, with simulated values y and v
  2. If $v < f_X(x)$, return $X = x$
  3. Otherwise go back to Step 1.
- Only requires simulating uniform random variables and evaluating the pdf pointwise



Rejection Sampling

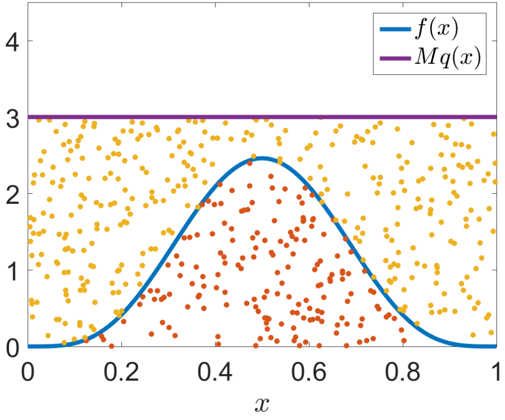
- Consider X a random variable on Ω with a pdf/pmf $f(x)$, a target distribution
- We want to sample from f using a proposal pdf/pmf q which we can sample.
- **Proposition**. Suppose we can find a constant M such that $f(x)/q(x) \leq M$ for all $x \in \Omega$.
- The following 'Rejection' algorithm returns $X \sim f$.

Algorithm 2 Rejection sampling

Step 1 - Simulate $Y \sim q$ and $U \sim \mathcal{U}[0,1]$, with simulated value y and u respectively.
Step 2 - If $u \leq f(y)/q(y)/M$ then stop and return $X = y$,
Step 3 - otherwise go back to Step 1.

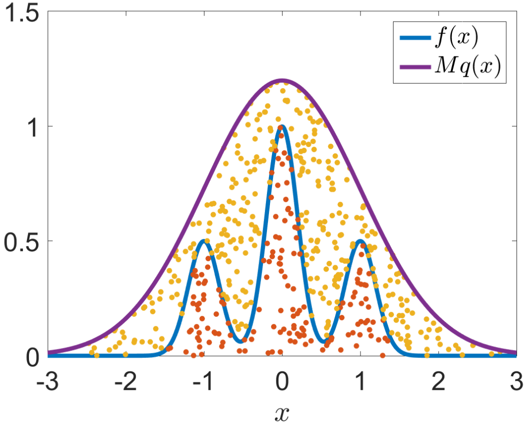
Illustrations

- $f(x)$ is the pdf of a $\mathrm{Beta}(5, 5)$ rv
- Proposal density q is the pdf of a uniform rv on $[0, 1]$



Illustrations

- $X \in \mathbb{R}$ with multimodal pdf
- Proposal density q is the pdf of a standardized normal



Rejection Sampling: Proof for discrete rv

- We have

$$\begin{aligned}\Pr(X=x) &= \sum_{n=1}^{\infty} \Pr(\text{reject } n-1 \text{ times, draw } Y=x \text{ and accept it}) \\ &= \sum_{n=1}^{\infty} \Pr(\text{reject } Y)^{n-1} \Pr(\text{draw } Y=x \text{ and accept it})\end{aligned}$$

- We have

$$\begin{aligned}&\Pr(\text{draw } Y=x \text{ and accept it}) \\ &= \Pr(\text{draw } Y=x)\Pr(\text{accept } Y \mid Y=x) \\ &= q(x)\Pr\left(U \leq \frac{f(Y)}{q(Y)}/M \,\middle|\, Y=x\right) \\ &= \frac{f(x)}{M}\end{aligned}$$

- The probability of having a rejection is

$$\begin{aligned}\Pr(\text{reject } Y) &= \sum_{x\in\Omega} \Pr(\text{draw } Y = x \text{ and reject it}) \\ &= \sum_{x\in\Omega} q(x) \Pr\left(U \geq \frac{f(Y)}{q(Y)}/M \,\middle|\, Y = x\right) \\ &= \sum_{x\in\Omega} q(x)\left(1 - \frac{f(x)}{q(x)M}\right) = 1 - \frac{1}{M}\end{aligned}$$

- Hence we have

$$\begin{aligned}\Pr(X = x) &= \sum_{n=1}^{\infty} \Pr(\text{reject } Y)^{n-1} \Pr(\text{draw } Y = x \text{ and accept it}) \\ &= \sum_{n=1}^{\infty} \left(1 - \frac{1}{M}\right)^{n-1} \frac{f(x)}{M} = f(x).\end{aligned}$$

- Note the number of accept/reject trials has a geometric distribution of success probability $1/M$, so the mean number of trials is M.

Rejection Sampling: Proof for continuous scalar rv

- Here is an alternative proof given for a continuous scalar variable X, the rejection algorithm still works but f, q are now pdfs.
- We accept the proposal Y whenever $(U, Y) \sim f_{U,Y}$ where $f_{U,Y}(u, y) = q(y)\mathbb{I}_{(0,1)}(u)$ satisfies $U \leq f(Y)/(Mq(Y))$.
- We have

$$
\begin{aligned}
\Pr(X \leq x) &= \Pr(Y \leq x | U \leq f(Y)/Mq(Y)) \\
&= \frac{\Pr(Y \leq x, U \leq f(Y)/Mq(Y))}{\Pr(U \leq f(Y)/Mq(Y))} \\
&= \frac{\int_{-\infty}^{x} \int_0^{f(y)/Mq(y)} f_{U,Y}(u, y) du dy}{\int_{-\infty}^{\infty} \int_0^{f(y)/Mq(y)} f_{U,Y}(u, y) du dy} \\
&= \frac{\int_{-\infty}^{x} \int_0^{f(y)/Mq(y)} q(y) du dy}{\int_{-\infty}^{\infty} \int_0^{f(y)/Mq(y)} q(y) du dy} = \int_{-\infty}^{x} f(y) dy.
\end{aligned}
$$

Example: Beta Density

- Assume you have for $\alpha, \beta \geq 1$

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1},\ 0 < x < 1$$

  which is upper bounded on $[0,1]$.
- We propose to use as a proposal $q(x) = \mathbb{I}_{(0,1)}(x)$ the uniform density on $[0,1]$.
- We need to find a bound M s.t. $f(x)/Mq(x) = f(x)/M \leq 1$. The smallest M is $M = \max_{0<x<1} f(x)$ and we obtain by solving for $f'(x) = 0$

$$M = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \underbrace{\left(\frac{\alpha-1}{\alpha+\beta-2}\right)^{\alpha-1} \left(\frac{\beta-1}{\alpha+\beta-2}\right)^{\beta-1}}_{M'}$$

  which gives

$$\frac{f(y)}{Mq(y)} = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{M'}.$$

Dealing with Unknown Normalising Constants

- In most practical scenarios, we only know $f(x)$ and $q(x)$ up to some normalising constants; i.e.

$$f(x) = \tilde{f}(x)/Z_f \text{ and } q(x) = \tilde{q}(x)/Z_q$$

  where $\tilde{f}(x), \tilde{q}(x)$ are known but $Z_f = \int_\Omega \tilde{f}(x)dx$, $Z_q = \int_\Omega \tilde{q}(x)dx$ are unknown/expensive to compute.
- Rejection can still be used: Indeed $f(x)/q(x) \leq M$ for all $x \in \Omega$ iff $\tilde{f}(x)/\tilde{q}(x) \leq \tilde{M}$, with $\tilde{M} = Z_f M/Z_q$.
- Practically, this means we can ignore the normalising constants from the start: if we can find $\tilde{M}$ to bound $\tilde{f}(x)/\tilde{q}(x)$ then it is correct to accept with probability $\tilde{f}(x)/\tilde{M}\tilde{q}(x)$ in the rejection algorithm. In this case the mean number N of accept/reject trials will equal $Z_q\tilde{M}/Z_f$ (that is, M again).

Simulating Gamma Random Variables

- We want to simulate a random variable $X \sim$Gamma(α, β) which works for any $\alpha \geq 1$ (not just integers);

$$f(x) = \frac{x^{\alpha-1}\exp(-\beta x)}{Z_f} \text{ for } x > 0, \quad Z_f = \Gamma(\alpha)/\beta^\alpha$$

  so $\tilde{f}(x) = x^{\alpha-1}\exp(-\beta x)$ will do as our unnormalised target.
- When $\alpha = a$ is a positive integer we can simulate $X \sim$ Gamma(a, β) by adding a independent Exp(β) variables, $Y_i \sim$ Exp(β), $X = \sum_{i=1}^{a} Y_i$.
- Hence we can sample densities 'close' in shape to Gamma(α, β) since we can sample Gamma$(\lfloor\alpha\rfloor, \beta)$. Perhaps this, or something like it, would make an envelope/proposal density?

- Let $a = \lfloor \alpha \rfloor$ and let's try to use Gamma(a, b) as the envelope, so $Y \sim$ Gamma(a, b) for integer $a \geq 1$ and some $b > 0$. The density of Y is

$$q(x) = \frac{x^{a-1}\exp(-bx)}{Z_q} \text{ for } x > 0, \quad Z_q = \Gamma(a)/b^a$$

so $\tilde{q}(x) = x^{a-1}\exp(-bx)$ will do as our unnormalised envelope function.

- We have to check whether the ratio $\tilde{f}(x)/\tilde{q}(x)$ is bounded over $\mathbb{R}_+$ where

$$\tilde{f}(x)/\tilde{q}(x) = x^{\alpha-a}\exp(-(\beta - b)x).$$

- Consider (a) $x \to 0$ and (b) $x \to \infty$. For (a) we need $a \leq \alpha$ so $a = \lfloor \alpha \rfloor$ is indeed fine. For (b) we need $b < \beta$ (not $b = \beta$ since we need the exponential to kill off the growth of $x^{\alpha-a}$).

- Given that we have chosen $a = \lfloor \alpha \rfloor$ and $b < \beta$ for the ratio to be bounded, we now compute the bound.
- $\frac{d}{dx}(\tilde{f}(x)/\tilde{q}(x)) = 0$ at $x = (\alpha - a)/(\beta - b)$ (and this must be a maximum at $x \geq 0$ under our conditions on a and b), so $\tilde{f}(x)/\tilde{q}(x) \leq \tilde{M}$ for all $x \geq 0$ if

$$\tilde{M} = \left(\frac{\alpha - a}{\beta - b}\right)^{\alpha - a} \exp(-(\alpha - a)).$$

- Accept Y at step 2 of Rejection Sampler if $U \leq \tilde{f}(Y)/\tilde{M}\tilde{q}(Y)$ where $\tilde{f}(Y)/\tilde{M}\tilde{q}(Y) = Y^{\alpha - a} \exp(-(\beta - b)Y)/\tilde{M}$.

Simulating Gamma Random Variables: Best choice of b

- Any $0 < b < \beta$ will do, but is there a best choice of b?
- Idea: choose b to minimize the expected number of simulations of Y per sample X output.
- Since the number N of trials is Geometric, with success probability $Z_f/(\tilde{M} Z_q)$, the expected number of trials is $\mathbb{E}(N) = Z_q \tilde{M}/Z_f$. Now $Z_f = \Gamma(\alpha)\beta^{-\alpha}$ where Γ is the Gamma function related to the factorial.
- Practice: Show that the optimal b solves $\frac{d}{db}(b^{-a}(\beta - b)^{-\alpha+a}) = 0$ so deduce that $b = \beta(a/\alpha)$ is the optimal choice.

Simulating Normal Random Variables

- Let $f(x) = (2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}x^2)$ and $q(x) = 1/\pi/(1+x^2)$. We have

$$\frac{\tilde{f}(x)}{\tilde{q}(x)} = (1+x^2)\exp\left(-\frac{1}{2}x^2\right) \leq 2/\sqrt{e} = \tilde{M}$$

  which is attained at ± 1.

- Hence the probability of acceptance is

$$\mathbb{P}\left(U \leq \frac{\tilde{f}(Y)}{\tilde{M}\tilde{q}(Y)}\right) = \frac{Z_f}{\tilde{M} Z_q} = \frac{\sqrt{2\pi}}{\frac{2}{\sqrt{e}}\pi} = \sqrt{\frac{e}{2\pi}} \approx 0.66$$

  and the mean number of trials to success is approximately $1/0.66 \approx 1.52$.

Rejection Sampling in High Dimension

- Consider

$$\tilde{f}(x_1, ..., x_d) = \exp\left(-\frac{1}{2}\sum_{k=1}^{d} x_k^2\right)$$

and

$$\tilde{q}(x_1, ..., x_d) = \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{d} x_k^2\right)$$

- For $\sigma > 1$, we have

$$\frac{\tilde{f}(x_1, ..., x_d)}{\tilde{q}(x_1, ..., x_d)} = \exp\left(-\frac{1}{2}\left(1-\sigma^{-2}\right)\sum_{k=1}^{d} x_k^2\right) \leq 1 = \tilde{M}.$$

- The acceptance probability of a proposal for $\sigma > 1$ is

$$\mathbb{P}\left(U \leq \frac{\tilde{f}(X_1, ..., X_d)}{\tilde{M}\tilde{q}(X_1, ..., X_d)}\right) = \frac{Z_f}{\tilde{M} Z_q} = \sigma^{-d}.$$

- The acceptance probability goes exponentially fast to zero with d.